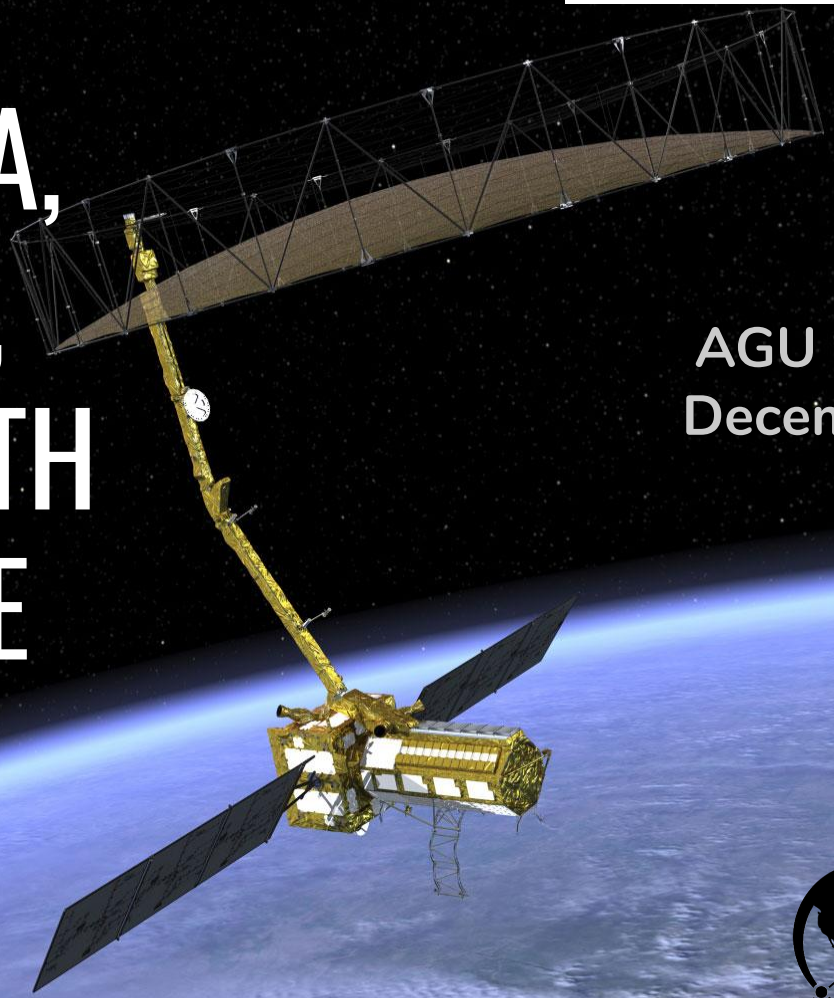
BIG DATA, CLOUD, AND EARTH SCIENCE



AGU Fall Meeting
December 9, 2019

TT12A

EARTHDATA
EOSDIS NASA'S EARTH OBSERVING SYSTEM DATA AND INFORMATION SYSTEM

Who am I?

Katie Baynes, katie.baynes@nasa.gov

System Architect,
NASA Earth Science Data and Information Systems

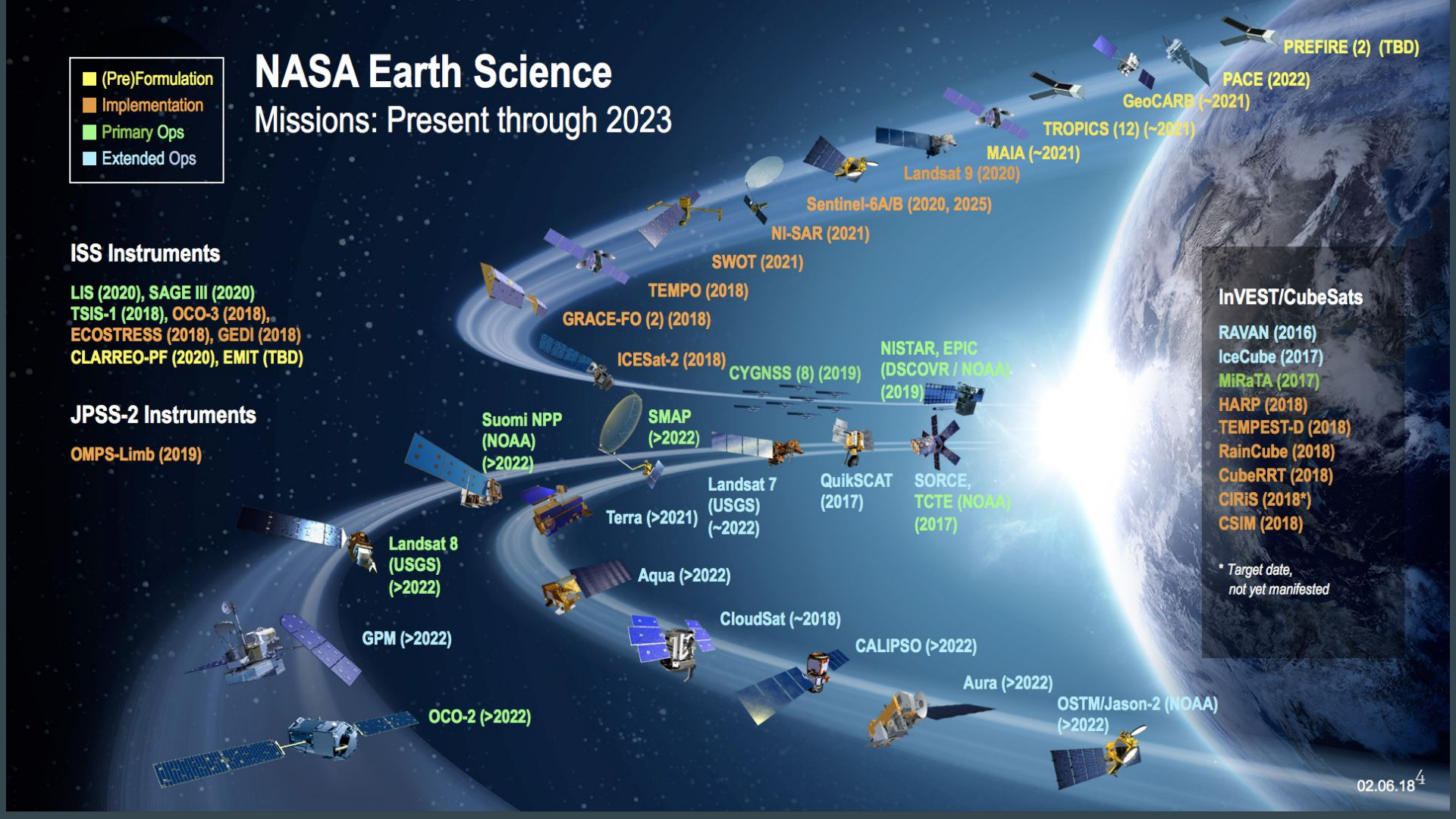
Currently working on migrating NASA Earth Science Data distribution to commercial cloud

NASA's
Earth Observing System Data and Information System

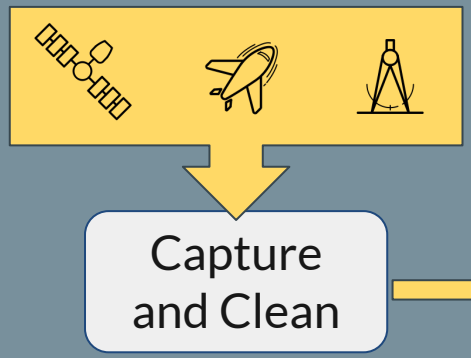
...

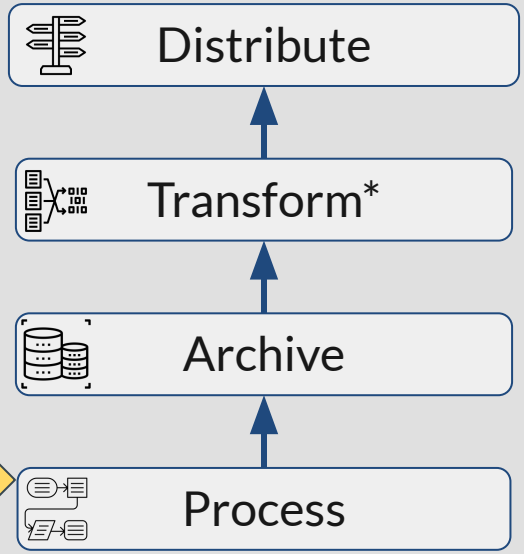
Our Work in Context

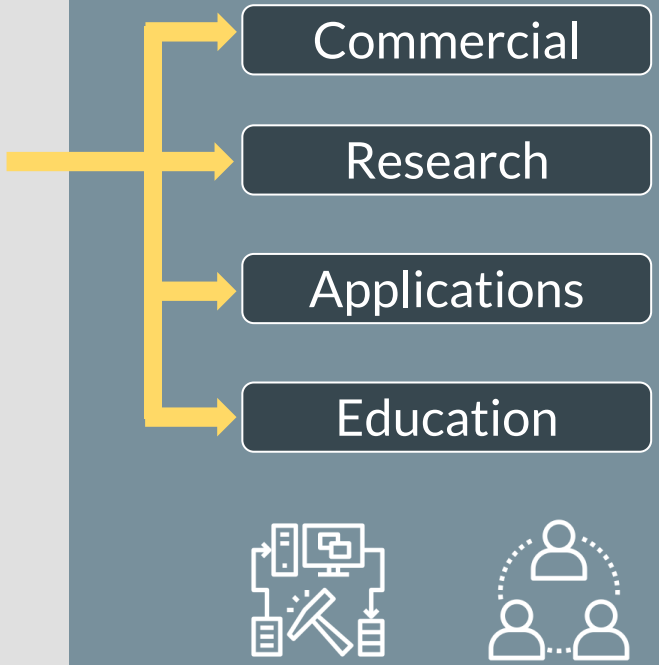
https://earthdata.nasa.gov


(Pre)Formulation
Implementation
Primary Ops
Extended Ops
NASA Earth Science
Missions: Present through 2023
ISS Instruments
LIS (2020), SAGE III (2020)
TSIS-1 (2018), OCO-3 (2018),
ECOSTRESS (2018), GEDI (2018)
CLARREO-PF (2020), EMIT (TBD)
JPSS-2 Instruments
OMPS-Limb (2019)
PREFIRE (2) (TBD)
PACE (2022)
GeoCARB (~2021)
TROPICS (12) (~2021)
MAIA (~2021)
Landsat 9 (2020)
Sentinel-6A/B (2020, 2025)
NI-SAR (2021)
SWOT (2021)
TEMPO (2018)
GRACE-FO (2) (2018)
ICESat-2 (2018)
CYGNSS (8) (2019)
NISTAR, EPIC
(DSCOVR / NOAA)
(2019)
Suomi NPP
(NOAA)
(>2022)
SMAP
(>2022)
Landsat 7
(USGS)
(~2022)
QuikSCAT
(2017)
SORCE,
TCTE (NOAA)
(2017)
Terra (>2021)
Landsat 8
(USGS)
(>2022)
Aqua (>2022)
CloudSat (~2018)
CALIPSO (>2022)
GPM (>2022)
Aura (>2022)
OCO-2 (>2022)
OSTM/Jason-2 (NOAA)
(>2022)
InVEST/CubeSats
RAVAN (2016)
IceCube (2017)
MiRaTA (2017)
HARP (2018)
TEMPEST-D (2018)
RainCube (2018)
CubeRRT (2018)
CIRiS (2018*)
CSIM (2018)
* Target date,
not yet manifested
02.06.18

EOSDIS in Context

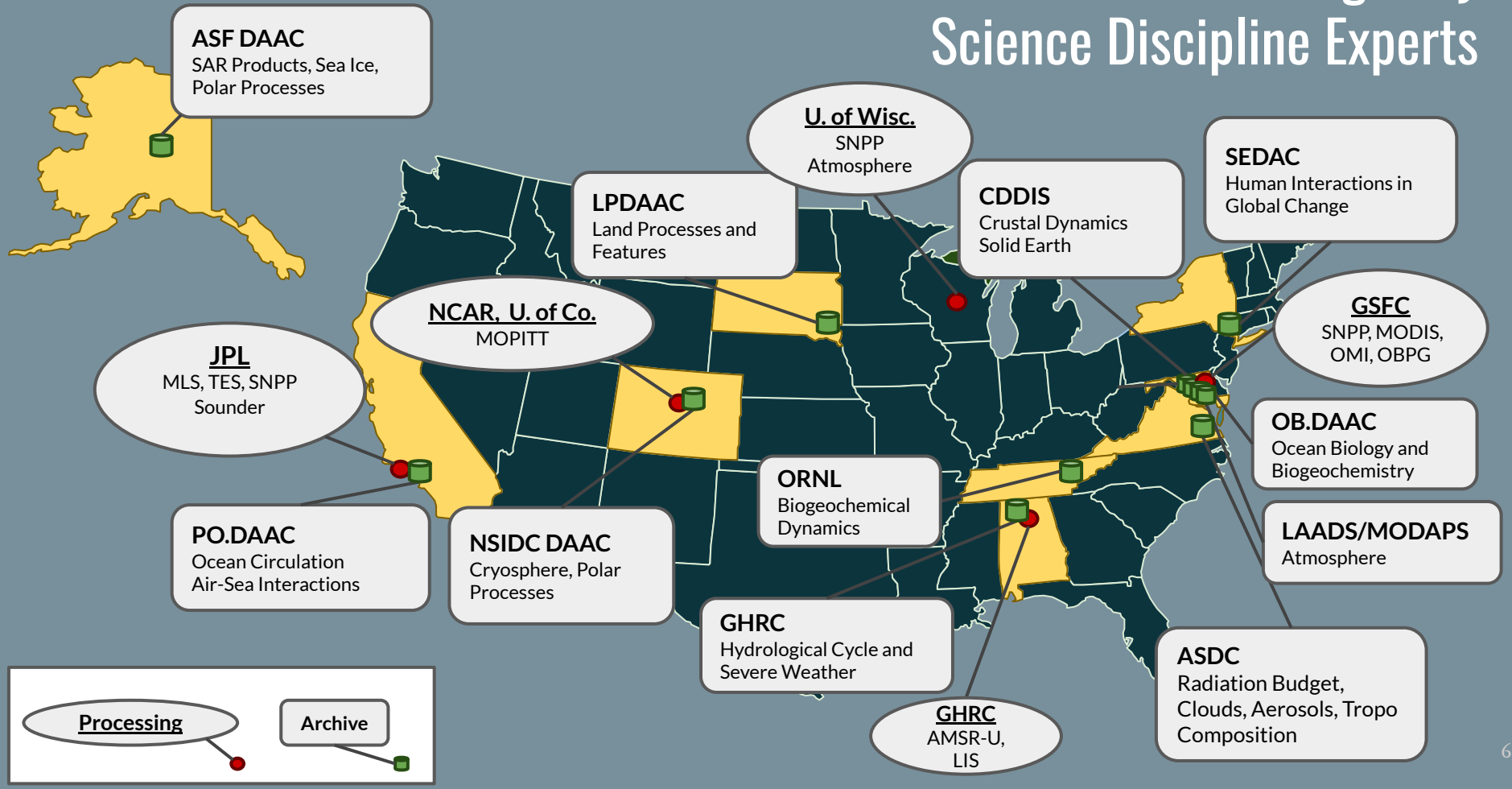






*Subset, reformat, reproject

Data are Produced and Managed by Science Discipline Experts



Big Variety at EOSDIS

NASA Earthdata Datasets in 2019

- ➔ **12** NASA centers of domain expertise
- ➔ **8,900** distinct data collections online
- ➔ **420** million cataloged files

Photo by Nathan Dumlao on Unsplash

EOSDIS Date Holdings Evolution

Data Look-ahead

Now
~ 23 TBs/day generated

Soon
~126 TBs/day generated

NASA's
Earth Observing System Data and Information System

...

Our Goals and Motivations

https://earthdata.nasa.gov

- ➔ Provide scientific **data stewardship** for all data collections and insure data integrity
- ➔ Provide a **unified and simplified environment** for a diverse and distributed community of Earth Science and Applications users
- ➔ **Evolve**, **grow and adapt** to new sources of data and new data systems technologies
- ➔ Expand the user community and engage with users to **enhance and improve user access** to data and other resources.
- ➔ **Partner** with other organizations, US agencies, and Nations to share data and make it easier to integrate for science

NASA's
Earth Observing System Data and Information System

...

Our Vision

https://earthdata.nasa.gov

EOSDIS Current Architecture

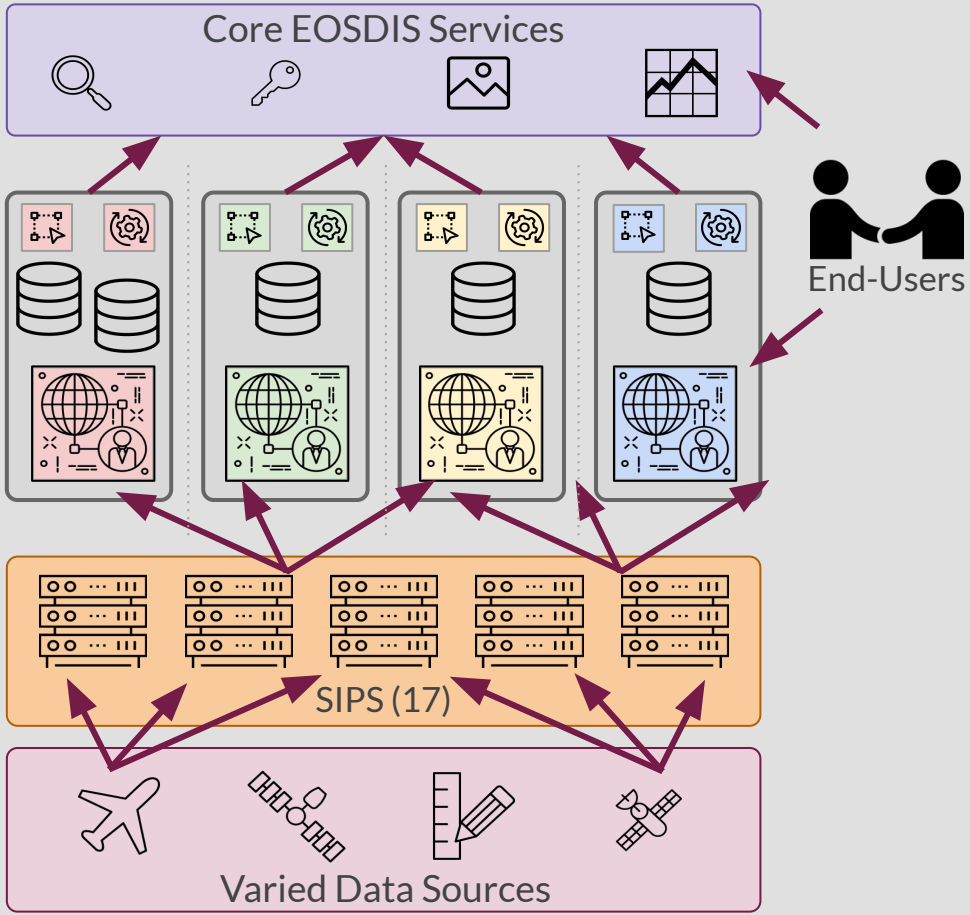
Benefits

Optimized for archive, search and distribution

Expert user support

Easily add new data products and producers

Predictable



Challenges

Uneven levels of service and performance

Significant time to coordinate interfaces

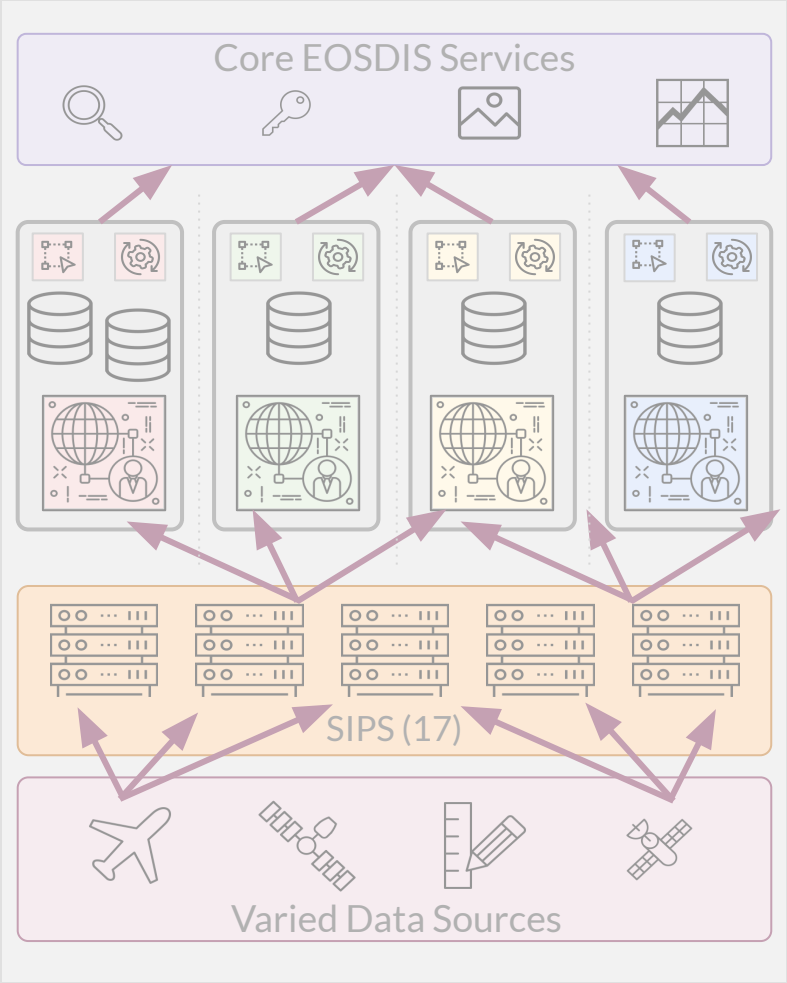
Limited on-demand product generation and end-user processing capabilities

Duplication of storage

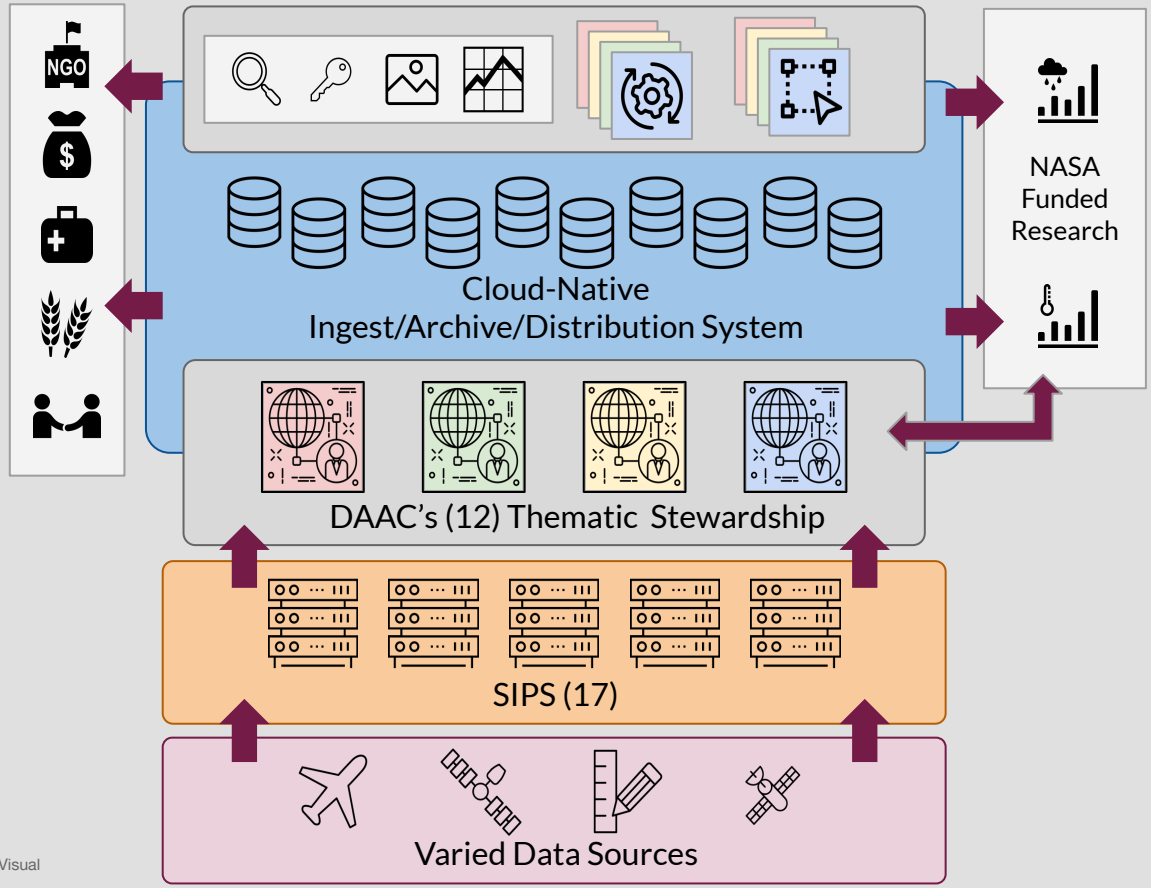
Duplication of services and software

We are on the cusp of opportunity. Can we do better? We are targeting:

- Better support for interdisciplinary Earth science researchers
- Reduced burden of data management/preparation for end-users
- More insightful, interactive data for research and commercial development
- More seamless interoperability with other institutional, international, and commercial providers
- Reducing overall monetary footprint and increasing efficiency across the system



Towards a Streamlined Cloud-Based Architecture



Benefits

Collocated, pay-as you-go processing for *anyone*

All data available to DAACs and users

Expert user support

Streamlined product addition

Challenges

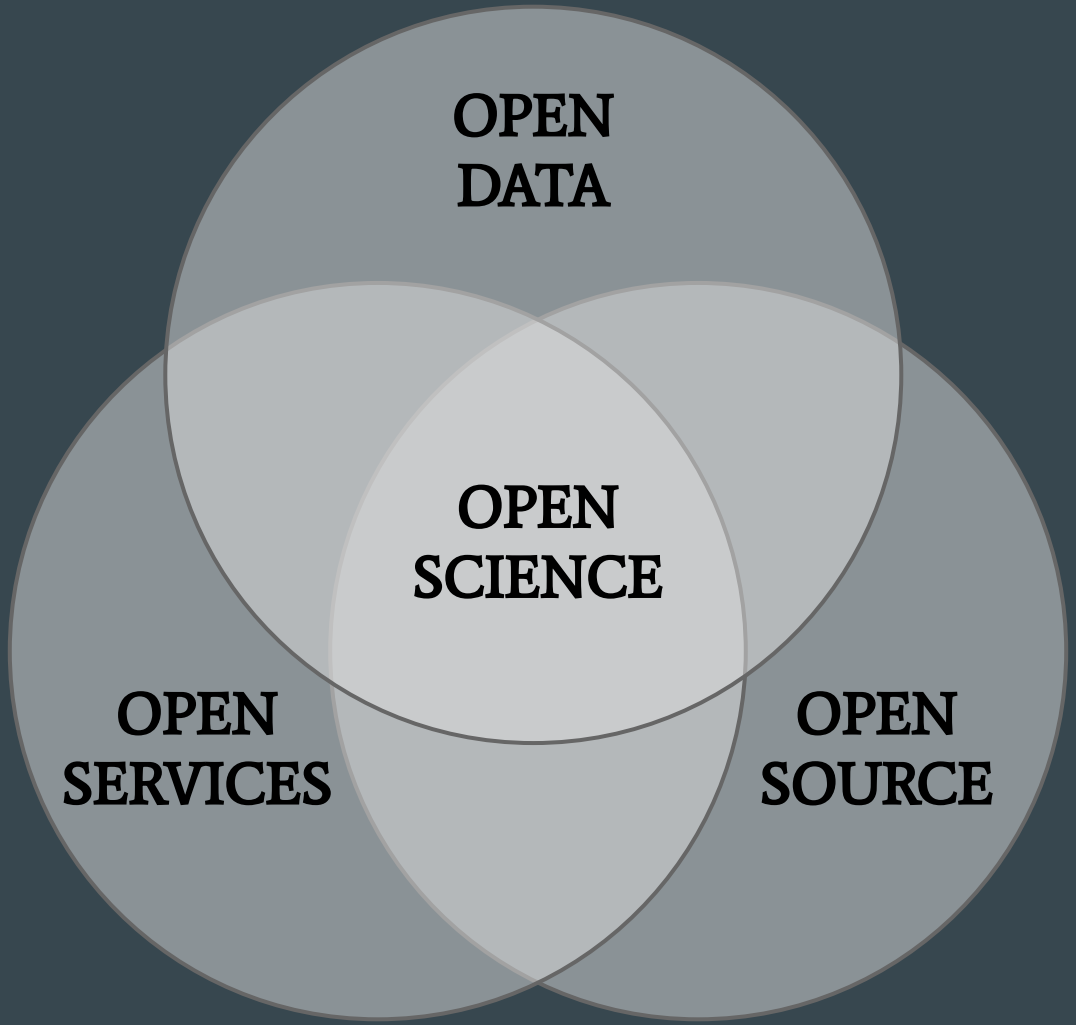
Development coordination

Cost Management

Shifting Labor Needs

Security/Export Compliance

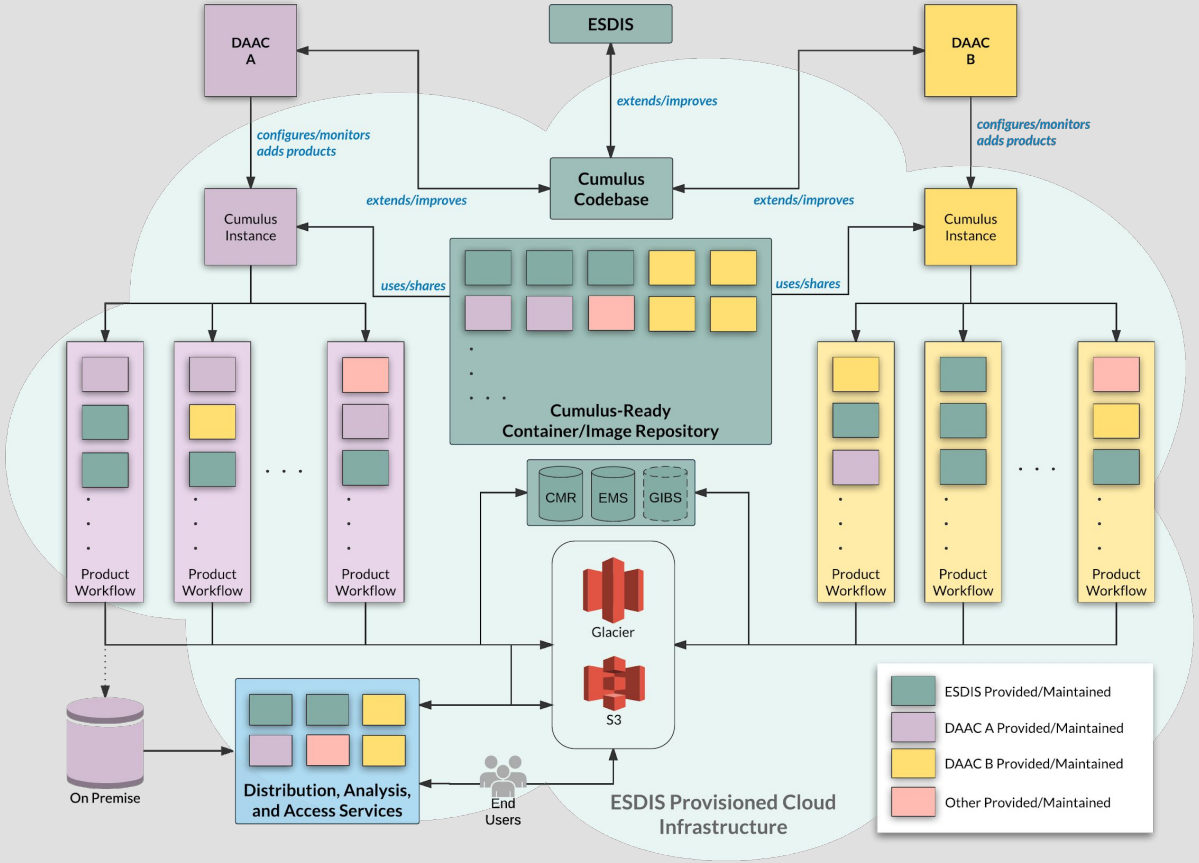
Vendor Lock In


OPEN
DATA
OPEN
SCIENCE
OPEN
SERVICES
OPEN
SOURCE

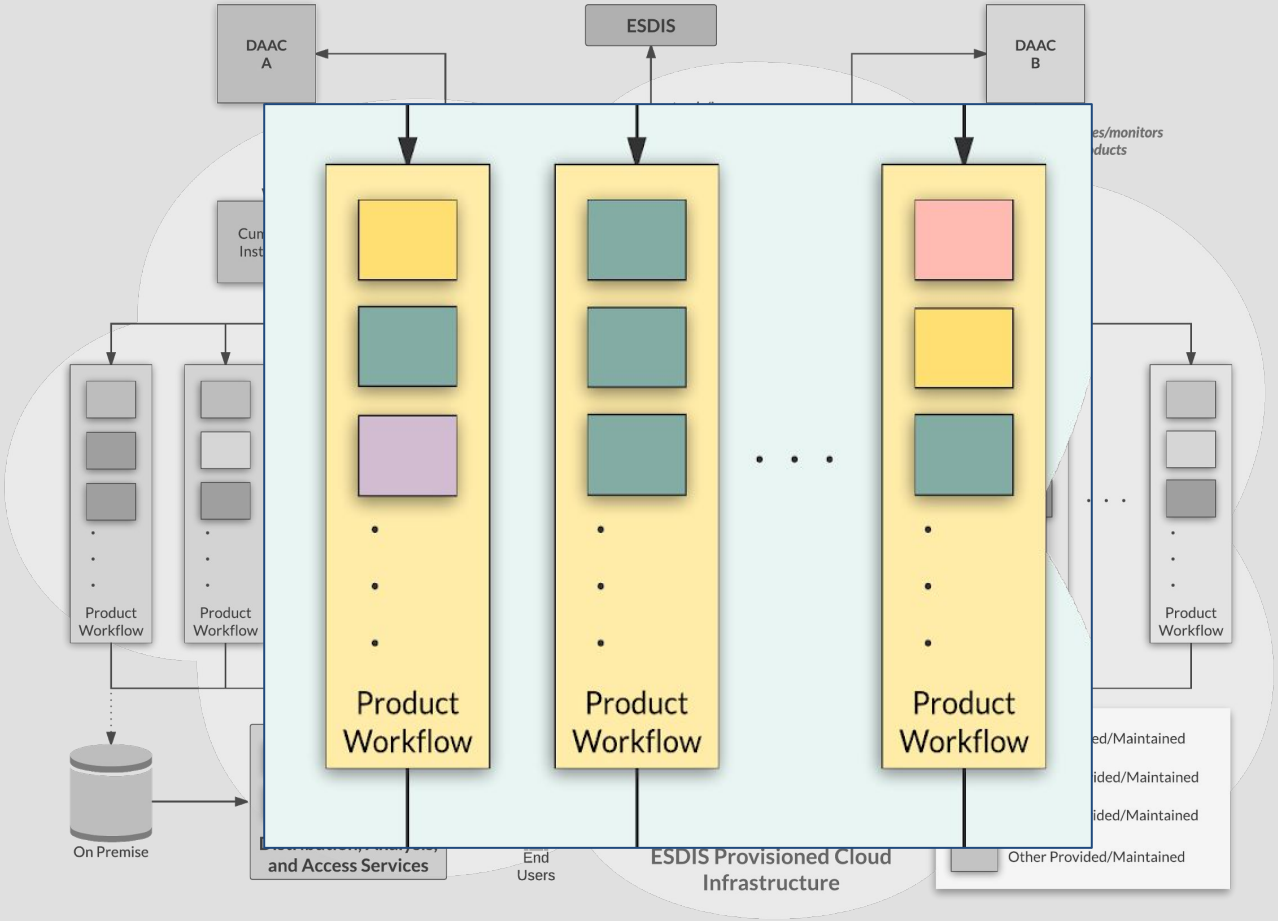
Unifying Ingest and Archive in the Cloud: Cumulus

EOSDIS DAACs all operate and maintain their own archive and distribution systems. This will also be how we operate in the future. However, as we work towards a cloud-based system, Cumulus is providing DAACs with a customizable

https://earthdata.nasa.gov

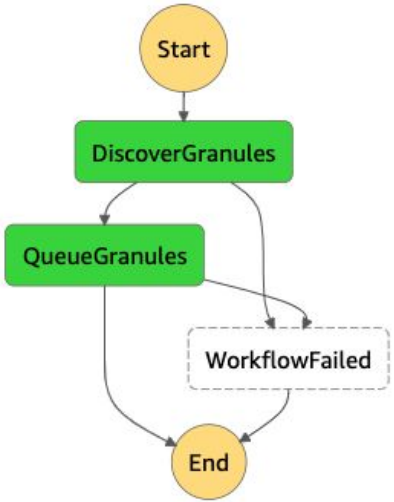


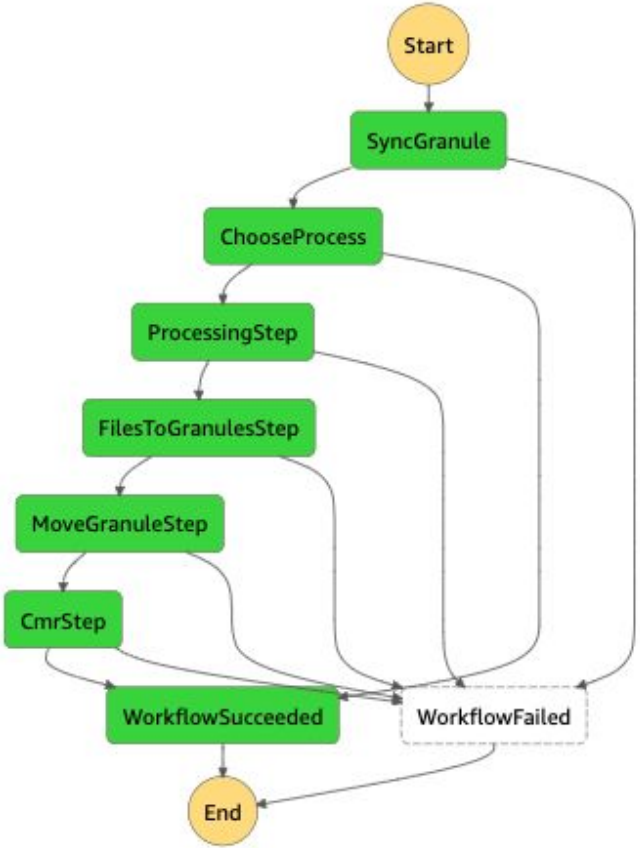
https://github.com/nasa/cumulus



Each DAAC maintains their own instances of Cumulus and other services.

Account owners have autonomy within their own AWS account.
(more on this later)





Example Step Function Cloud Workflows

https://aws.amazon.com/step-functions/

Unifying Services in the Cloud: Harmony

Historically, EOSDIS DAACs have all provided their own tooling with diverse interaction patterns and APIs. Harmony is our ongoing effort to revisit these siloed capabilities in a more harmonized manner.

https://earthdata.nasa.gov

The Harmony Elevator Pitch

End-Users

Service Execution Framework(s)

Enterprise Integration: Login, Metrics, Egress, Metadata Catalog

Common Interface: Common API, Earthdata Search UI

Data staging, Transformation code execution

DAAC- or Core-Team-Supplied

DAAC-Unique Transformations

e.g., SWOT water feature averaging

DAAC-Supplied

Common Transformations

e.g., Subsetting L3 NetCDF

DAAC- or Core-Team-Supplied

NASA's
Earth Observing System Data and Information System

...

Our Current Status

https://earthdata.nasa.gov

Current EOSDIS Systems Operating in the AWS Cloud

Common Metadata Repository

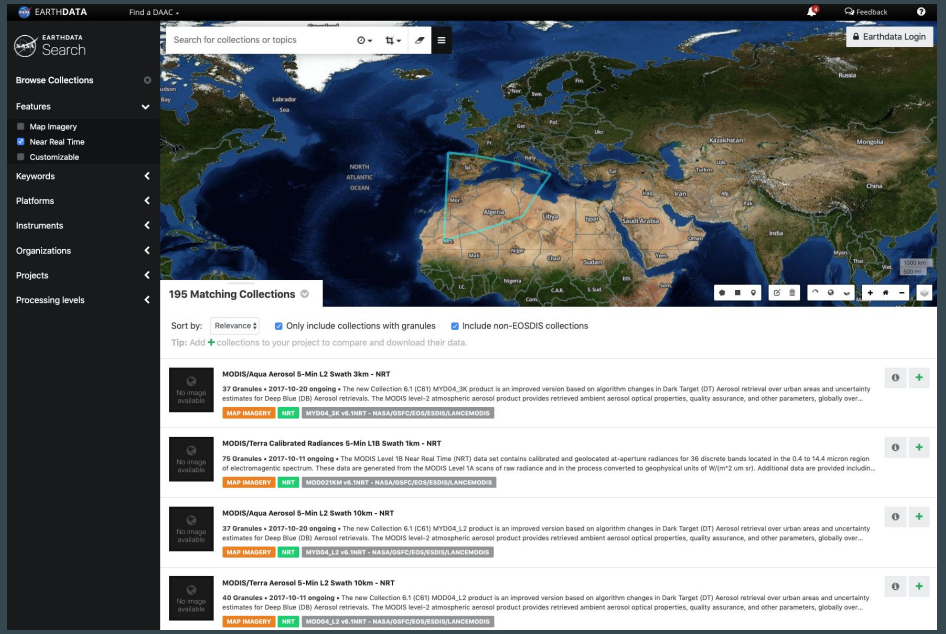
https://cmr.earthdata.nasa.gov

Earthdata Search

https://search.earthdata.nasa.gov

API-driven, standards-compliant,
sub-second search of:

- 8,900 collections
- 420 million files



Current EOSDIS and Partner Data in the AWS Cloud

Global Hydrology Research Center

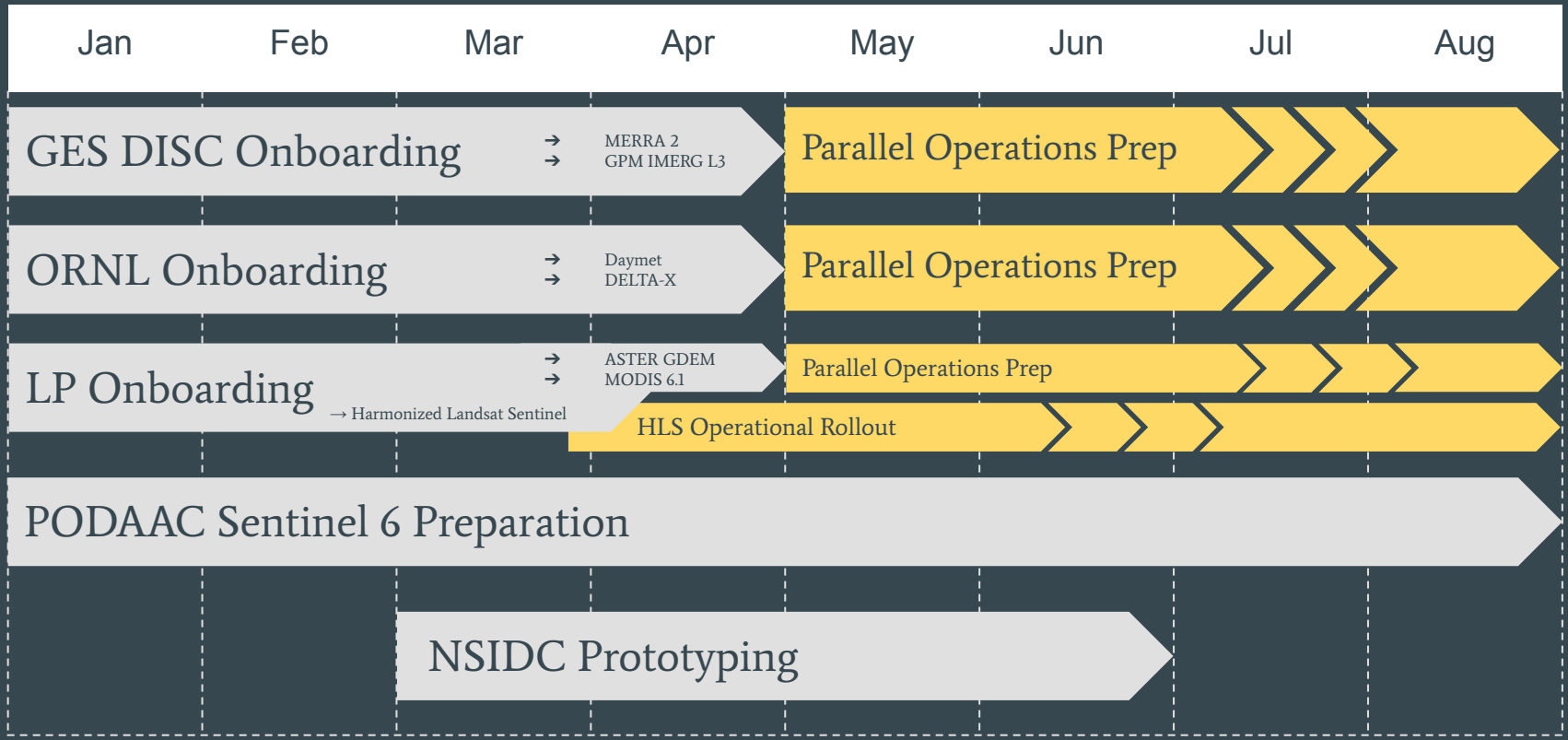
https://ghrc.nsstc.nasa.gov/home/

Alaska Satellite Facility

ESA's Sentinel 1 Archive Mirror
https://search.asf.alaska.edu/

https://media.asf.alaska.edu/uploads/home-cards/satellite-dish-scenic.jpg

Planned Cloud Dataset Timeline in 2020



NASA's
Earth Observing System Data and Information System

...

Our Challenges and Strategies

https://earthdata.nasa.gov

Challenge:
Vendor Lock-In

Photo by Cristina Gottardi on Unsplash

Data Transfer Risk

“What if you have to move the data?”

Right now, AWS is the only NASA-approved commercial cloud vendor. As more options become available we will investigate them.

Application Transfer Risk

"<XYZ> AWS-specific product!"

Most of the tools we are using are not a unique problem that Amazon alone has solved.

There are usually (many) free and open source, alternatives.

As we continue to evolve cloud functionality, we continue to examine trade-offs between out-of-the-box and vendor agnostic.

https://unsplash.com/photos/IcSoX12oagU

Infrastructure Transfer Risk

What about ECS, Lambda, SQS, etc

Again, these are not unique problems. Every major competitor in the cloud space has alternatives, or open source alternatives exist.

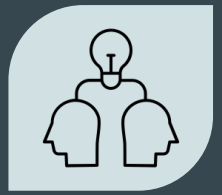
Serverless: Qinling, Google Cloud Functions

Queues: Zaqar, RabbitMQ

etc, etc

https://unsplash.com/photos/Oaqk7qqNh_c

Knowledge Transfer Risk



"We are training everyone in AWS"

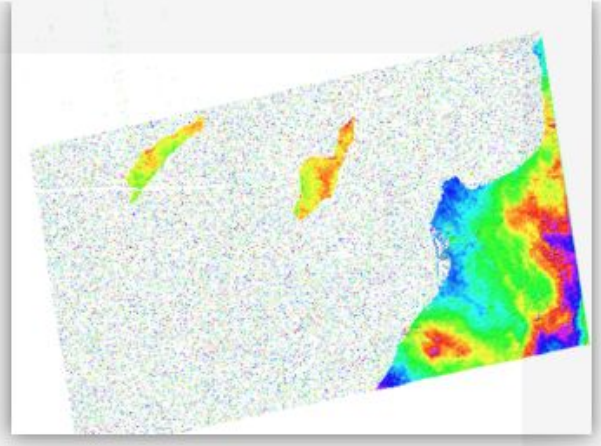
This is a real problem. Effectively leveraging the AWS console is its own skillset. People may become unwilling to be retrained if we have to migrate. But we have faced this problem before.

Challenge:
End-User Adoption

If the data is in the cloud, we would like to encourage users to work with that data in place.

https://www.nasa.gov/sites/default/files/thumbnails/image/icebridge_201705_arctic2.jpg

All DAACs are starting to look at this problem individually…



Data Recipes

How to Create and Unwrap an Interferogram with GMT5SAR Script in the Cloud — Windows

In this document you will find:

We are Developing a “Cloud Primer” to aid user transition across all DAACs

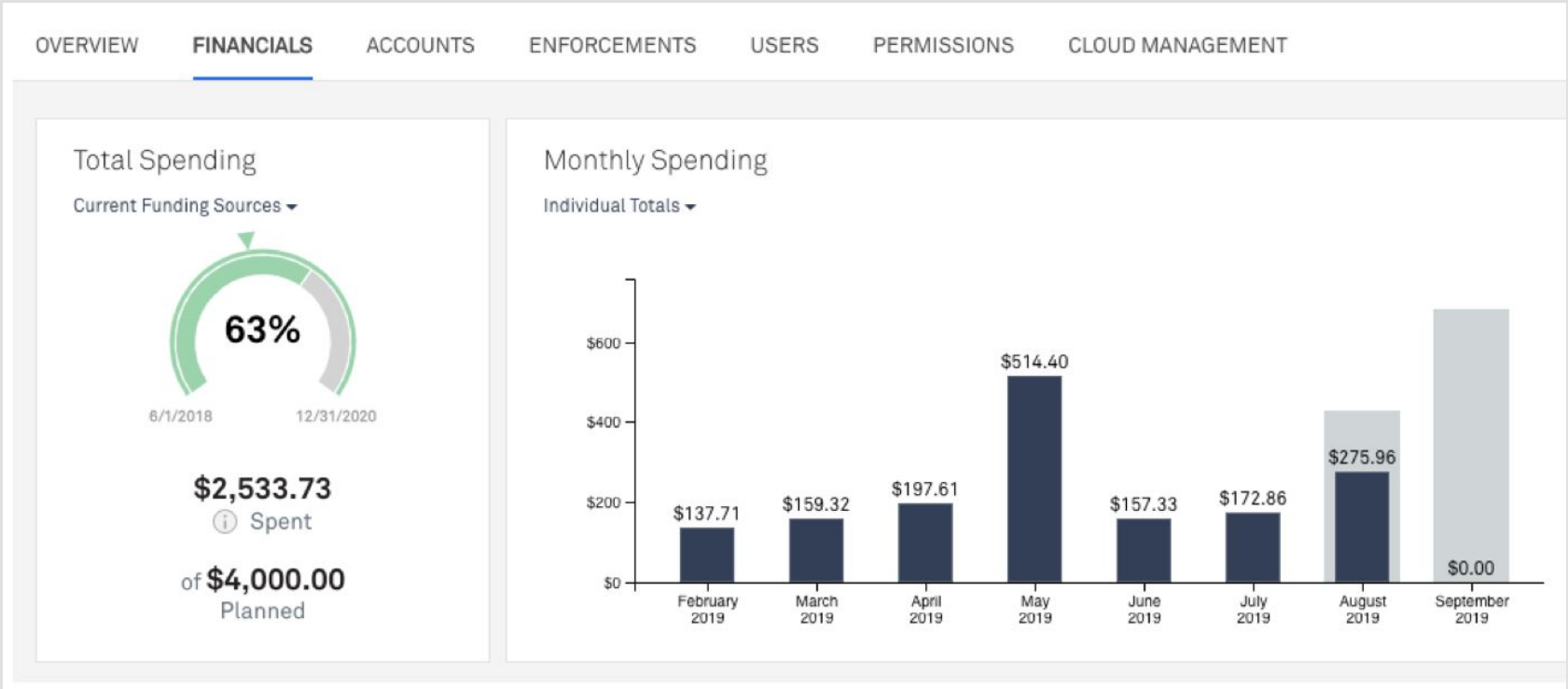
POC:
justin.rice@nasa.gov

https://pixabay.com/en/help-climbing-hand-mentor-2444110/

Challenge:
Cost Control

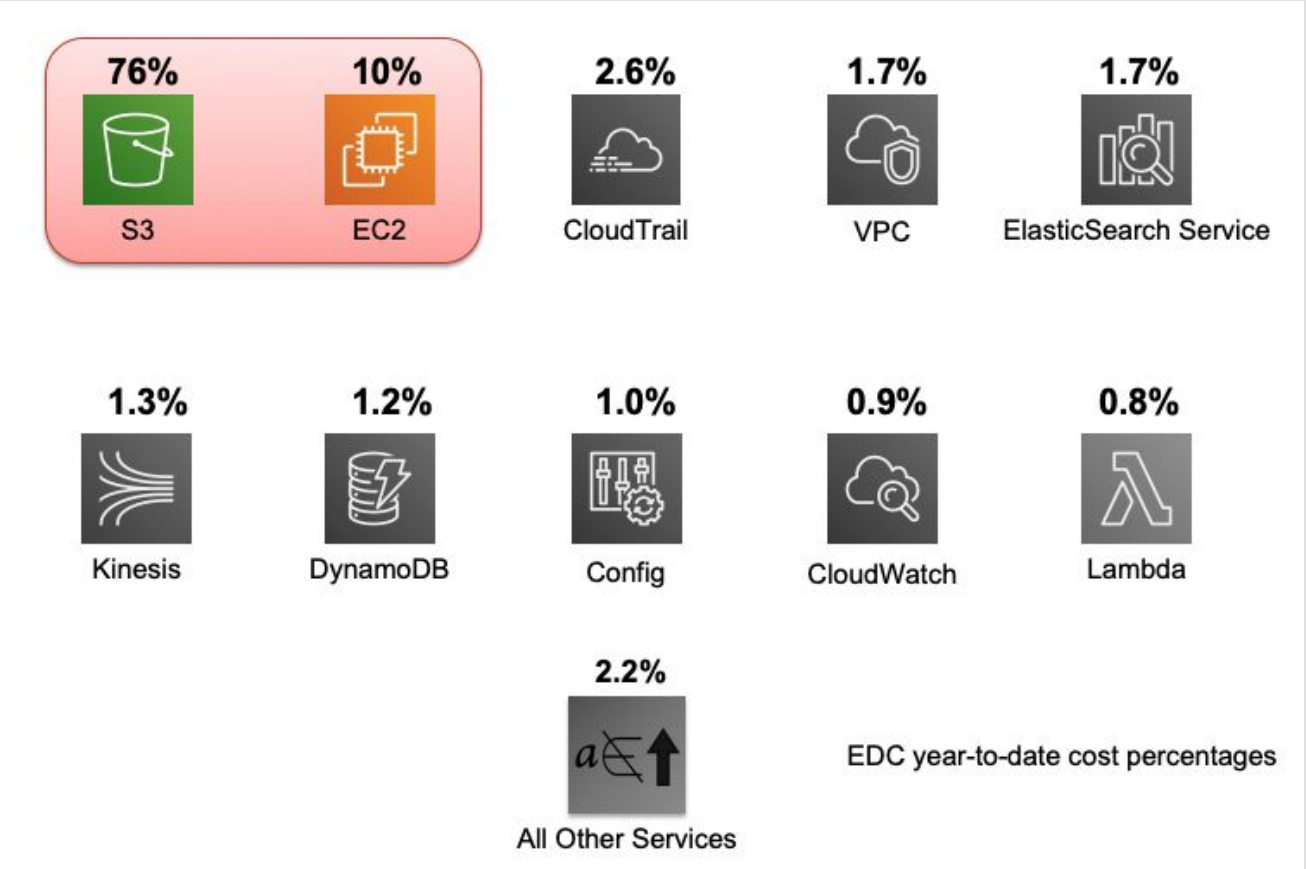
Photo by Fabian Blank on Unsplash

Budget and Cost Monitoring

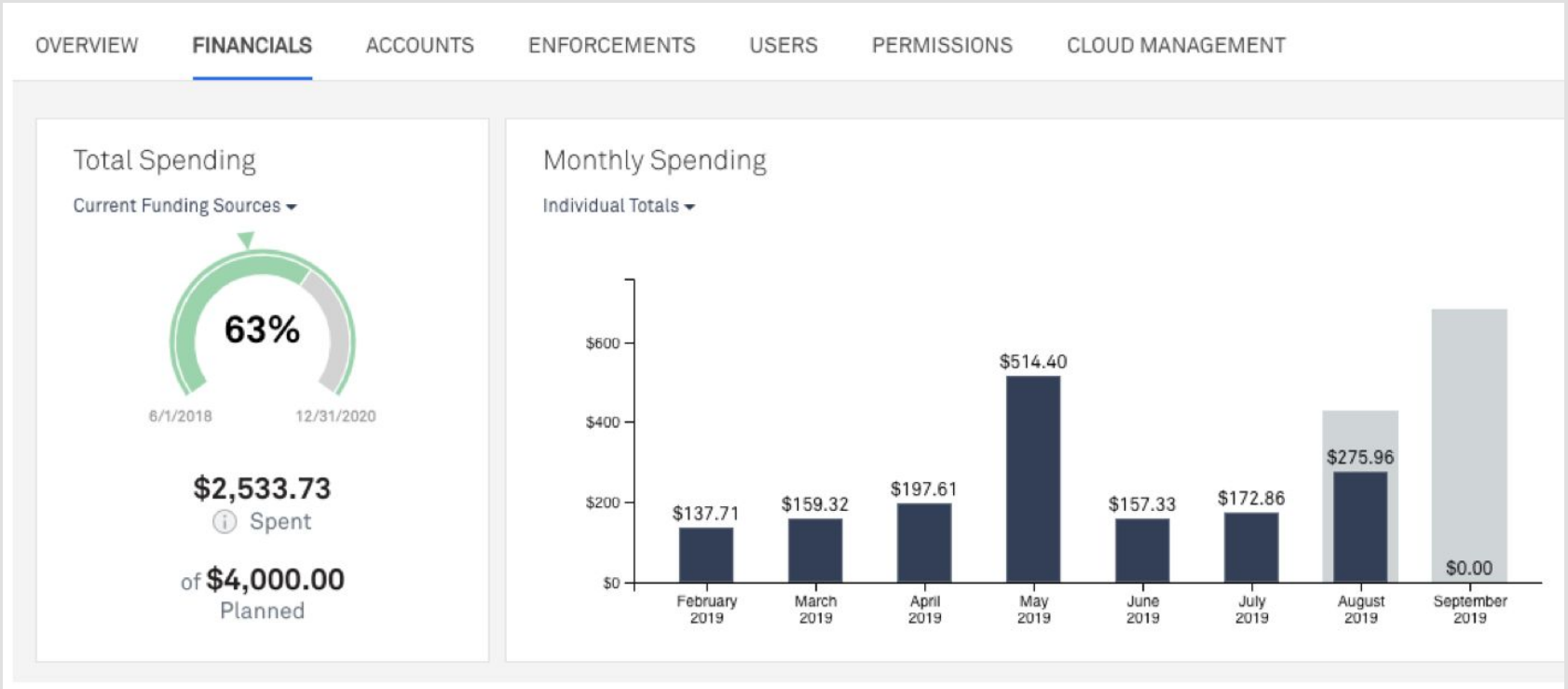


https://www.cloudtamer.io/

Cost Conscious Development



Budget and Cost Monitoring



https://www.cloudtamer.io/

Challenge: Security

IT Security keeps us safe; keeping up-to-date and protected requires constant vigilance

Code Security Working Group

Security is every team member's responsibility. We are working towards automating as much as possible to remain productive and safe as we migrate to the cloud.

We regularly test and implement new tools to protect our code bases, dependency trees, and operational systems and vulnerabilities.

- Code Security Working Group PMB Sub-Team
  - 2018-09-25 Meeting notes
  - 2018-10-03 Meeting notes
  - 2019-03-27 Meeting notes
  - 2019-05-29 Meeting notes
  - **2019-10-16 Meeting notes**
  - Information: Current processes for deploying code
  - Proposal: Defense in depth for programmatic secrets

Limiting Exposure while Communicating with Users

Lambda

Full Access w/ Caveats

All Lambda functions must execute from within the NGAP provisioned private subnets of the Application VPC. . .

Lambda Networking Requirements	
VPC	**Application VPC**
Subnet	**Private application-[xxx]**
Security Group	**Any**
IAM Helper Policy	**NGAPShLambdaInVpcBasePolicy** Contains all necessary permissions for a generic Lambda to execute from within the Application VPC. This policy should be attached, along with any other required policies, to your Lambda IAM execution role.

IAM

Identity and Access Management

Limited Permissions

Application owners may *only* manage IAM permissions for application components running in their AWS account(s), not for users. All IAM roles created by app owners are subject to NGAP-managed permissions boundaries.

IAM Requirements	
IAM Roles	**NGAPShRoleBoundary / NGAPShNonProdRoleBoundary . . .** **MUST** be assigned as Permissions Boundary to create a custom IAM Role.
IAM Policies	**Full Access**
IAM Users	Handled via CloudTamer. No access through AWS IAM Web Console.
IAM Access Keys	Handled via CloudTamer. No access through AWS IAM Web Console.

Other Opportunities for Big Data in the Cloud

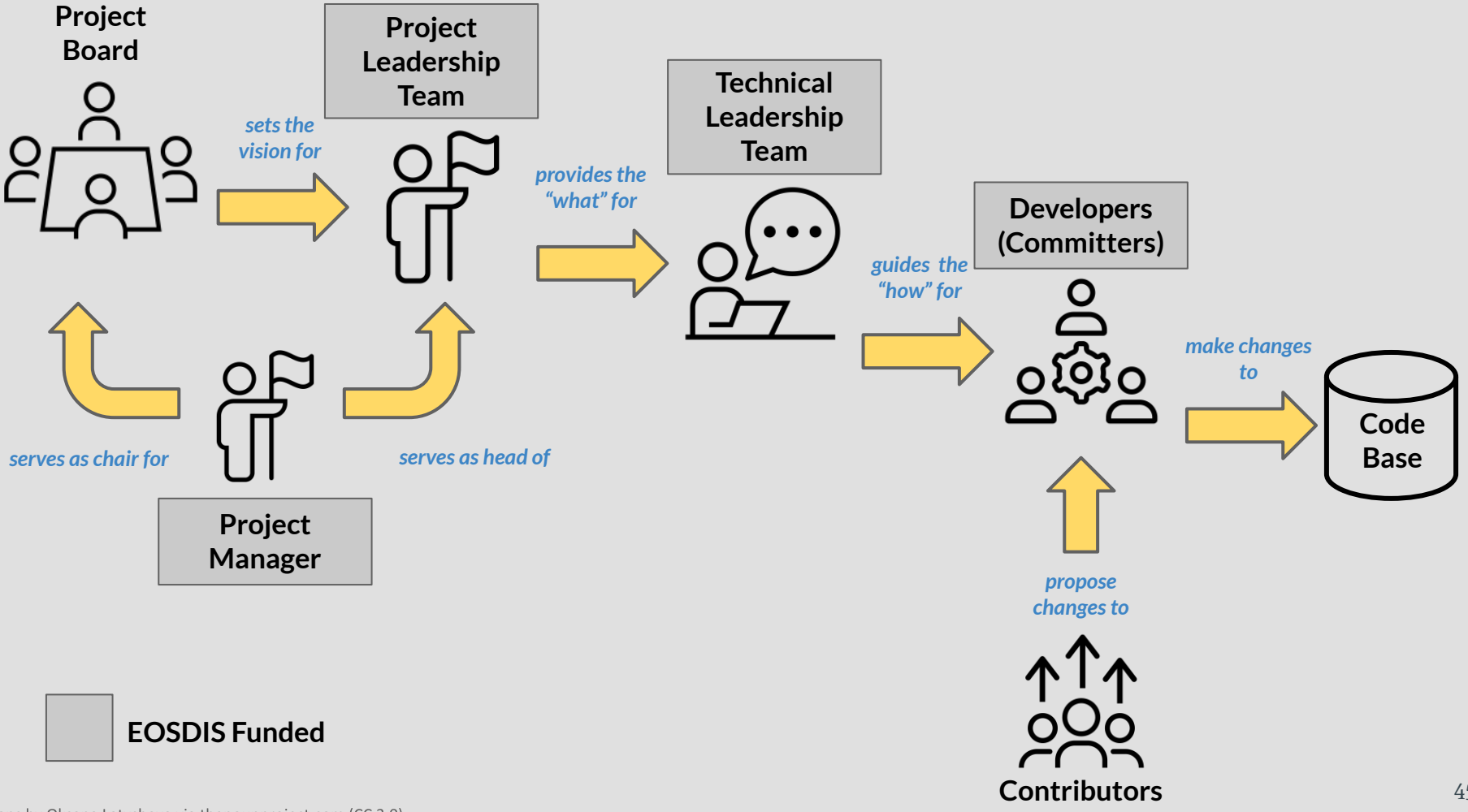
Much of our work over the last 24-36 months has been about coping with a oncoming data onslaught. We still have much ground to tread to embrace our new found wealth.

https://earthdata.nasa.gov

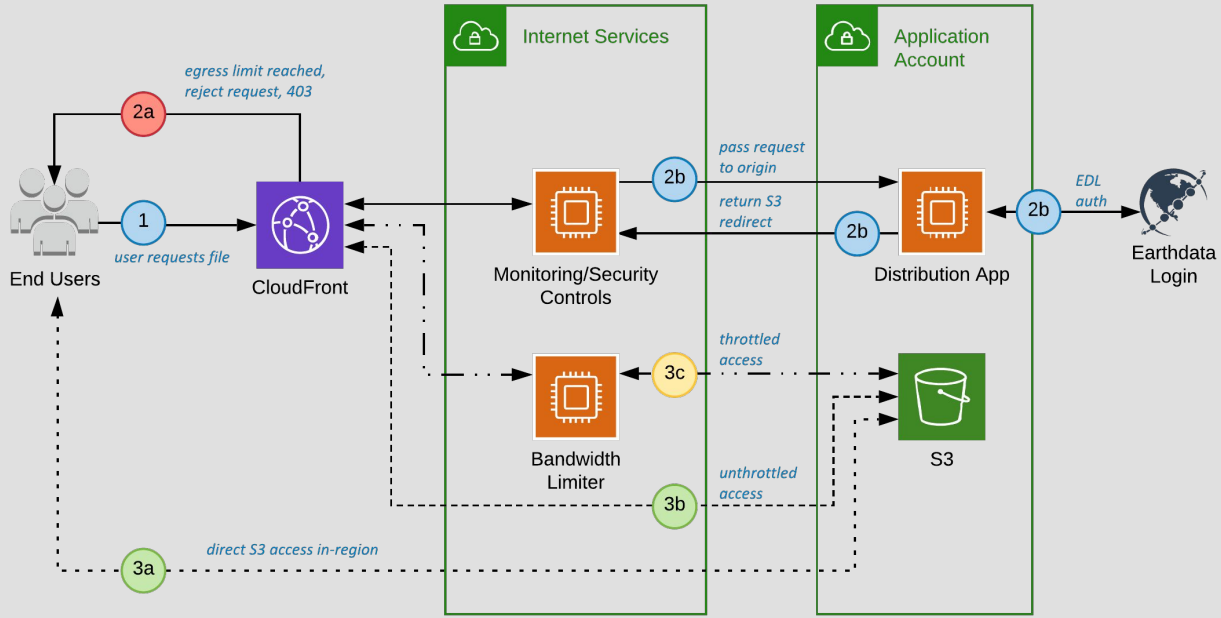
Questions?

https://earthdata.nasa.gov

Backups



Tackling Egress Monitoring



1 DNS points to NGAP-controlled CloudFront distribution.

2a If the egress cutoff limit for this month has already been reached, the request is rejected (403).

2b Otherwise, the request is passed to the origin (a tenant app/distribution app, e.g. Cumulus), through the platform's monitoring stack (Internet Sevices). If frequired, the application's distribution application sends unauthenticated users and those with expired sessions to Earthdata Login for Auth. A signed S3 URL is returned.

If the origin returns a redirect to S3, CloudFront then picks between the following download mechanisms:

3a If the application user is in the **same region** as the S3 bucket, pass through the S3 redirect unchanged; the user downloads directly from S3.

3b If the application user is *not* in the same region but throttling is turned **off**, the user is redirected to a CloudFront URL to download the data (unthrottled) from S3 w/ CloudFront cost savings in place.

3c If the application user is not in the same region but throttling is turned ***on***, the download from S3 will be proxied through the bandwidth limiter.

Otherwise, the response will be returned to the application user unmodified.